# Macroeconomic forecasting through news, emotions and narrative

Sonja Tilly[a,*], Markus Ebner[b] and Giacomo Livan[a,c]

[a]*UCL, Computer Science Dep, 66 - 72 Gower St, Bloomsbury, WC1E 6EA London, UK*

[b]*Quoniam Asset Management, Westhafen Tower, Westhafenplatz 1, 60327 Frankfurt am Main, Germany*

[c]*Systemic Risk Centre, London School of Economics and Political Science, London, WC2A 2AE, UK*

## ARTICLE INFO

## ABSTRACT

This study forecasts industrial production and consumer prices leveraging narrative and sentiment from global newspapers. Existing research includes positive and negative tone only to improve macroeconomic forecasts, focusing predominantly on large economies such as the US. These works use mainly anglophone sources of narrative, thus not capturing the entire complexity of the multitude of emotions contained in global news articles. This study expands the existing body of research by incorporating a wide array of emotions from newspapers around the world – extracted from the Global Database of Events, Language and Tone (GDELT) [32] – into macroeconomic forecasts. We present a thematic data filtering methodology based on a bi-directional long short term memory neural network (Bi-LSTM) for extracting emotion scores from GDELT and demonstrate its effectiveness by comparing results for filtered and unfiltered data. We model industrial production and consumer prices across a diverse range of economies using an autoregressive framework, and find that including emotions from global newspapers significantly improves forecasts compared to an autoregressive benchmark model. We complement our forecasts with an interpretability analysis on distinct groups of emotions and find that emotions associated with surprise and happiness have the strongest predictive power for the variables we predict.

## 1. Introduction

Recent developments in automated language analysis have allowed to quantify the elusive yet intuitive notion of narrative, and to quantify its predictive power in relation to changes in social systems.

Research in psychology and cognitive sciences has examined the role emotions and narrative play in decision making and judgement [8, 9, 14]. These studies show that emotions can help individuals make decisions in complex scenarios with uncertain outcomes. Keynes uses the term "animal spirits" to describe the dispositions and emotions that drive human actions, with the results of this behaviour measurable in terms of economic indices such as consumer confidence [29]. Shiller [46] finds that unsettling narrative led to events such as the Great Depression in the 1920s and the Global Financial Crisis in 2008/9, arguing that narrative is a means of predicting the economy. A recent theoretical development – known as Conviction Narrative Theory (CNT) – draws on the concept that to be sufficiently confident to act, agents create narratives supporting their expectations of the outcome of their actions [40]. For instance, a study on CNT tracks changes in narrative and shows that they precede changes in economic growth [52].

Media is an established, multi-functional tool for governments, corporations and individuals to disseminate information, connect and interact. As such, it is a major conduit for news narrative. Nowadays, most forms of media have an online presence and produce huge volumes of data. This data contains information in the form of opinions and sentiment about financial markets and the economy, which may not yet be reflected in macroeconomic variables.

Over recent years, researchers have explored sentiment from different types of media and its usefulness for the prediction of the economy. Studies examine how to process large amounts of unstructured data from a variety of sources in order to extract signals [10, 19]. Other works outline approaches to incorporate such signals into a predictive model, for instance to improve the monitoring of the economy and financial forecasting [35, 47].

Media sentiment prediction has a wide range of application domains that Rousidis et al [43] group into finance, marketing and sociopolitical. Within the finance domain, studies explore media sentiment prediction either for specific assets or markets (micro level) [1] or for different aspects of the economy (macro level) [2].

Existing research incorporates mainly positive and negative tone to improve macroeconomic forecasts, thus not capturing the entire complexity of the multitude of emotions contained in global news articles. Most works use anglophone sources of narrative, focusing predominantly on large economies such as the US.

This study advances the existing body of research by incorporating a wide array of emotions from newspapers around the world into macroeconomic forecasts using data from the Global Database of Events, Language and Tone (GDELT) [32]. GDELT is a research collaboration that analyses global news articles and extracts items such as themes, emotions, locations, and many more. We employ a filtering methodology based on machine learning to identify articles that are relevant to the macroeconomic indices in question, and provide a proof of concept demonstrating that emotions expressed in those news items add value to forecasts of industrial production and consumer prices across a diverse range of economies, both in terms of geographic location and size. We complement this with dimensionality reduction and corre-

lation analysis in order to group the more than 600 emotion scores available into a smaller number of interpretable factors. We find emotions associated with "surprise" and "happiness" to yield the highest predictive power across the variables we forecast. To the best of our knowledge, emotions from GDELT Global Knowledge Graph's Content Analysis Measure Systems have not yet been used to forecast macroeconomic variables.

## 2. Literature review

This section addresses a selection of existing literature on macroeconomic forecasting with media sentiment.

A rapidly evolving body of literature examines the use of media sentiment and big data for economic forecasting [11, 28, 48]. The majority of studies forecast economic variables with regression frameworks combining traditional data with positive and negative sentiment classifications based on word count (as opposed to a wider spectrum of emotions).

Research suggests that positive and negative sentiment from newspaper narrative is an effective tool for monitoring the economic cycle [46, 52]. Similarly, newspaper narrative is found to precede a change in economic variables with low frequency shifts correlating well with financial market events. Hence, newspaper narrative can be regarded as a risk management tool [40].

While most studies focus on a single economy, Baker et al [3, 4] develop indices of economic uncertainty for a wide range of countries. They use an autoregressive framework including variables derived from news as well as macroeconomic variables to gauge whether uncertainty shocks foreshadow weaker macroeconomic performance. Findings suggest that effects of policy uncertainty on firms and macro data raises stock price volatility, lowers investment rates and employment growth. Political bias does not significantly impact the uncertainty indices. Thorsrud [51] decomposes unstructured newspaper text into daily news topics and uses them to forecast quarterly GDP growth, producing significantly better predictions compared to central bank forecasts. A study by Pekar and Binner [41] demonstrates that adding information on intended purchases from Twitter tweets alongside lagged consumer index values often yields statistically significant improvements over the baseline model that is trained with lag variables alone.

Newspaper archives and Twitter are commonly used sources for raw textual data, however there is a growing body of research using preprocessed sentiment scores. Ortiz [12] combines official statistics with themes from GDELT to track Chinese economic vulnerability in real-time, showing that the index provides valuable insights for policymakers and investors. Elshendy et al [20] combine data from GDELT with a set of traditional macroeconomic variables and use social network analysis to generate predictors for macroeconomic indices such as consumer confidence, business confidence and GDP for the 10 largest EU economies. Results show that data extracted from GDELT is valuable for predicting macroeconomic variables. Chen [13] examines the effect of the negative narrative in relation to international trade from

US presidential candidates in 2016 using average tone from GDELT. The study concludes that narrative can impact the economy by influencing market participants' expectations. Glaeser et al [21] use reviews from YELP to forecast the local economy. Results from a regression analysis suggest that the data set is a useful complement for predicting contemporaneous changes in the local economy. It also provides an up-to-date snapshot of economic change at local level, delivering the best results for populous areas and the hospitality industry, given the high number of reviews.

While most publications argue in favour of using media sentiment for macroeconomic forecasting, Schaer et al [45] take a more critical view, highlighting the need for thorough statistical testing, careful choice of error metrics and benchmarks. They acknowledge some of the challenges when using sentiment data such as data complexity, sampling instability and key word selection.

The majority of literature only incorporates positive and negative tone to improve macroeconomic predictions, with just a handful of studies featuring a wider range of emotions in their analyses. This paper expands the existing body of research by incorporating nuanced sentiment from newspapers around the world into macroeconomic forecasts of industrial production and consumer prices for 10 diverse economies. This paper goes beyond mere prediction and also focuses on the interpretability of results, illustrating which emotions have the strongest predictive power.

## 3. Data and methods

This section introduces GDELT as data source, outlines the filtering methodology that is used and provides information about the nature of the sentiment scores.

The GDELT Project is a research collaboration of Google Ideas, Google Cloud, Google and Google News, the Yahoo! Fellowship at Georgetown University, BBC Monitoring, the National Academies Keck Futures Program, Reed Elsevier's LexisNexis Group, JSTOR, DTIC, and the Internet Archive. The project monitors world media from a multitude of perspectives, identifying and extracting items such as themes, emotions, locations and events. GDELT version two incorporates real-time translation from 65 languages and measures over 2,300 emotions and themes from every news article, updated every 15 minutes [32]. It is a public data set available on the Google Cloud Platform.

The Global Knowledge Graph (GKG), one of the tables within GDELT, contains fields such as sentiment scores and themes extracted from global newspaper articles. It comprises around 11 terabytes of data with new data being added constantly, starting in February 2015. To date, it has analysed over one billion news items.

### 3.1. Predicted variables

This study models industrial production (IP) and consumer price indices (CPI) for US, UK, Germany, Norway, Poland, Turkey, Japan, South Korea, Brazil and Mexico. IP is a monthly measure of economic activity. It is defined as the output of industrial establishments, covers a broad

range of sectors and tracks the change in the volume of production output. The consumer price index (CPI) is selected as monthly inflation index and describes the change in the prices of a basket of goods and services that are typically purchased by households.

### 3.2. Filtering methodology

A filtering methodology is applied to extract sentiment scores from GDELT's GKG relevant to economic growth and inflation, respectively, containing three steps:

- Step 1: Keyword filter

- Step 2: Classification with neural network

- Step 3: Aggregation

Step one consists of a top-level thematic filter based on keywords (economic growth, inflation) to select relevant articles based on themes. Step two uses a neural network to further filter news items using GDELT themes. Step three aggregates the sentiment scores to the frequency of the macroeconomic variables. In addition, this step applies country filters to GDELT locations.

To filter out non-relevant information, a simple keyword filter is applied to GKG themes. The GDELT algorithm extracts themes from every news article it analyses [34]. The GKG contains over 12,000 unique themes.

An analysis of a random set of 100 original news articles is conducted to evaluate the keyword filter's ability to eliminate non-relevant news items, showing that the GDELT algorithm has a tendency of recognising themes where there are none. This creates the need to further filter observations using the themes column. For each news article scanned, the themes are a sequence of labels, with each label representing a theme. Given the nature of the data, this is a text sequence classification problem. A set of 1,000 random articles is manually classified according to relevance. This is done by looking up the original news articles using the DocumentIdentifier column, which corresponds to the article's url. In cases where the url is no longer available, the news item is disregarded. The themes for every article are label-encoded so that the themes are given numbers between zero and $N - 1$, where $N$ is the total number of themes. For out-of-vocabulary words, an "unknown" token is assigned.

A range of classifiers are trained on 800 observations and tested out of sample on 200 observations, where the encoded themes are the predictor variables and the classifications into relevant/non-relevant articles are the predicted variables. Performance is assessed using the area under the curve (AUC) score. He and Ma [25] suggest that the AUC is a more appropriate metric for the assessment of a classifier than basic accuracy, especially in the case of imbalanced data, as the latter is too biased towards the dominant class. Table one shows the performance of different algorithms that were evaluated on a data set filtered for economic growth.

| Classifier | AUC |
|---|---|
| Gaussian NaÃŕve Bayes | 0.65 |
| Random Forest | 0.67 |
| Support Vector Machine | 0.75 |
| XGBoost | 0.88 |
| Unidirectional NN | 0.82 |
| Bi-LSTM | 0.95 |

Table 1: Classifier performance

After evaluating a range of algorithms, a Bi-LSTM neural network is selected as it delivers the strongest performance. While recurrent neural networks have difficulties learning long-term dependencies, long short neural networks are able to preserve information from inputs that has already passed through its hidden state [26]. A Bi-LSTM architecture is well suited to tasks where context is important such as sequence classification. The algorithm runs inputs simultaneously into two directions – from the past to the future and from the future to the past – thus preserving information from both past and future at every step [23].

The filtered data from step 2 is aggregated according to time period and location filters are applied to GKG locations according to each country's economic links. The location column contains a list of all locations found in each news item, extracted through the algorithm designed by Leetaru [31].

In order to gain insights into the economic interconnectedness of each of the 10 countries, the import and export volumes by trading partner are examined [16]. Six of the economies have diversified trade links with countries around the world. Poland, Norway and Turkey trade predominantly with Western European economies. For South Korea, over half of the country's imports and exports are linked to China. Due to the trade links between economies, information relating to one country may also be relevant for another one [42]. Based on this idea of interconnection, a global data set incorporating information on all 10 countries is generated for the six global economies (US, UK, Germany, Japan, Brazil, Mexico). For Poland, Norway and Turkey, a data set containing information on Western European economies is generated. For South Korea, a data set including information on China is built.

### 3.3. Nature of sentiment scores

Within GDELT's GKG, the Tone and the Global Content Analysis Measures (GCAM) column contain over 2,300 sentiment scores.

The Tone field comprises a comma-delimited list of six emotional dimensions, each recorded as floating point number. From this field, the average tone of the document is used. This score typically ranges from -10 (very negative) to +10 (very positive), with zero being neutral [33]. The tone score is based on sentiment mining. This approach counts words according to positive and negative pre-compiled dictionaries. The net sentiment represents the overall tone [27].

The GCAM system runs 24 content analysis systems over each news article and returns the resulting scores as a comma-delimited list into the GCAM column [33]. The majority of GCAM scores is based on word count, some are based on more sophisticated methods. GCAM also includes the overall word count for each news item analysed.

There is some overlap between the GCAM scores generated by the different analysis systems. Scores of following four analysis systems are chosen as they minimise duplication of sentiment scores while incorporating a broad range of emotions:

- WordNet-Affect was developed by Strapparava and Valitutti [50] based on WordNet Domains [38]. WordNet Domains maps synsets, i.e. groupings of synonymous words expressing the same concept, to domain labels such as Economics or Health. WordNet Affect extends this structure in assigning affective domain labels to the synsets. WordNet-Affect scores are word count-based and account for 280 sentiment dimensions in the GCAM column.

- The Loughran and McDonald Financial Sentiment Dictionary uses negative word lists specific to a financial context to produce scores based on word count. The authors find that word lists for other disciplines often misclassify words in financial documents [37]. The system generates six scores.

- The Hedometer scores provide a measurement for overall societal happiness for English and a range of non-English languages. In order to provide an overall score, over 10,000 unique words are rated by humans on a scale from one to nine. For each of these words, an average happiness score is derived, with five being neutral. The system returns 12 scores.

- ML-Senticon represents a multi-layered synset-level lexicon and calculates positivity and negativity scores covering English and Spanish. This system first uses a number of algorithms to estimate the polarity of individual synsets [15]. Subsequently, an average happiness score for a news item is calculated [17]. The system generates 32 scores.

The extracted data is aggregated by month. The mean and standard deviation of the tone score is calculated. Where the GCAM sentiment scores are based on word count, the mean and standard deviation are calculated, normalized to account for variation of word count as done by Baker et al. [4]. For calculated sentiment scores, the mean score and standard deviation over the period are computed. In addition, the number of news items and the total word count per period is generated.

### 3.4. The data sets
This section sets out how the data sets used in this study are built.

The filtering methodology is applied to build two data sets, filtered for articles relevant to economic growth and inflation, respectively. In step one, country filters for the US, UK, Germany, Norway, Poland, Turkey, Japan, South Korea, Brazil and Mexico are applied. The countries are selected to represent a diverse mix of economies, both in size and geography. The data is aggregated to monthly frequency, from beginning of March 2015 to end of June 2020, respectively. Model predictions incorporate true positives, false positives, true negatives and false negatives. The filtered data sets comprise true positive and false positive predictions only. They include c 5.4% and c 3.9% of noise for the economic growth and the inflation filter, respectively.

An unfiltered data sample of aggregated GCAM scores is created for comparison purposes. Around five million random observations (one million for each calendar year) are selected and aggregated to monthly frequency. The unfiltered data set contains over 60% noise i.e. news items not relevant to economic growth or inflation, respectively.

In order to account for macroeconomic effects, the baltic dry index and the crude oil price are incorporated when modelling IP. The baltic dry index is a leading indicator for economic activity, reflecting levels of global trade [6]. A study by Van Eyden et al [53] suggests that there is a significant relationship between oil price fluctuations and economic growth in OECD countries. For models forecasting CPI, the countries' respective terms of trade indices as well as the crude oil price are included. Mihailov et al [39] find that the anticipated relative change in the terms of trade is a more important determinant of inflation than the contemporaneous domestic output gap. A study by Salisu et al establishes a significant long-term positive relationship between oil price and inflation [44].

### 3.5. Data preprocessing
In this section the data preparation methods are summarized.

For the each of the 10 aforementioned economies the respective values for IP and CPI are used as predicted variables. Both index values represent the monthly percentage change.

The augmented Dickey Fuller unit root test is applied to 20 years of monthly data and rejected at 5% significance for the above described variables.

Where a sentiment score contains zeros only, it is assumed that the relevant GCAM system did not return any scores and they are dropped from the respective data set. The scores affected are mainly based on the Hedometer and ML Senticon GCAM systems. The GDELT data sets contain 664 raw scores and this step reduces the amount of features to 630 and 628 for data sets filtered for economic growth and inflation, respectively. The unfiltered data set retains 632 features. The monthly change in sentiment scores is calculated. The augmented Dickey Fuller unit root test is applied and stationary is not rejected at 5% for any of the scores. The sentiment scores are standardized by removing the mean and

scaling to unit variance.

The most recent 20% of the data are set aside as "unseen" data for testing purposes.

## 4. Analysis

This section outlines the analysis that is performed to gauge if the sentiment scores from GDELT have predictive power.

### 4.1. Granger causality analysis

The Granger causality between the GDELT sentiment scores and the predicted variables is assessed as to evaluate if there are relationships between those variables. While Granger causality can provide useful insights into the relation between variables, it is not testing true causality, instead, the test looks to establish if changes in one variable occur before changes in the other one [22]. This means that Granger causality may be found even when there is no causal link [30].

The null hypothesis for the Granger causality test states that lagged sentiment scores are not causing a variable at a significance level of 5%, while the alternate hypothesis stipulates that lagged sentiment scores are Granger-causing an index at the same significance level.

The Granger causality for lags up to a maximum of three months is evaluated. Since multiple tests for each data set are run, the resulting *p*-values are adjusted according to the Benjamini-Hochberg (BH) procedure to control for multiple hypothesis testing [5].

### 4.2. Dimensionality reduction

Principal component analysis (PCA) is then applied to the GDELT sentiment scores to discount redundancies between these features due to correlations between them, which could lead to overfitting during modelling. PCA is widely used in literature to reduce high-dimensional data. Stock and Watson [49] show that a small number of principal components extracted from a large data set can be used to predict macroeconomic indices. Similarly, Hanson and McMahon apply principal component analysis to features extracted from FOMC communications and demonstrate that the principal components have an effect on economic variables [24].

PCA is applied to the data sets filtered for IP and CPI, respectively. The first eight principal components are used, explaining between 45% and 48% of the variance.
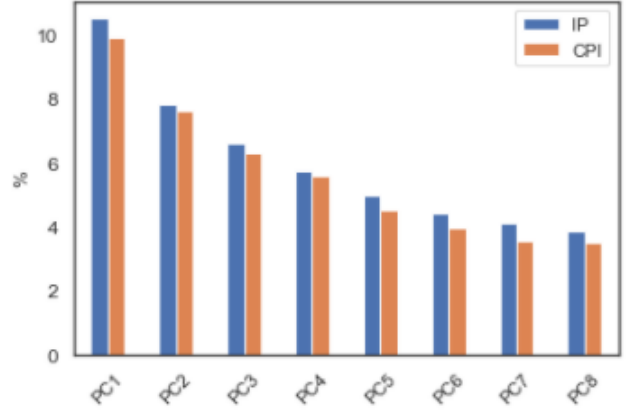


**Figure 1:** Variance explained by the first eight principal components (global data sets)

Figure 1 illustrates the percentage of variance each principal component explains extracted from the global data sets.

### 4.3. Forecasting

As a further step in the analysis of the sentiment scores, an autoregressive framework is used for forecasting macroeconomic variables.

This framework allows modeling a $T \times K$ multivariate time series $Y$, where $T$ denotes the number of observations and $K$ the number of variables. The framework is defined as

$$Y_t = v + A_1 Y_{t-1} + \cdots + A_p Y_{t-p} + u_t \qquad (1)$$

where $A_i$ is a $K \times K$ coefficient matrix, $v$ is a constant and $u_t$ is white noise. The model is then calibrated for each macroeconomic variable and each country.

The optimal lag length is selected based on the the Akaike (AIC) and the Bayesian (BIC) information critera. These measures are based on the idea that the inclusion of a further term may improve the model however the model should also be penalised for increasing the number of parameters to be estimated. When the improvement in goodness-of-fit outweighs the penalty term, the statistic associated with the information criterion decreases. Thus, the lag which minimises the information criterion is selected [7].

For each country, the respective country macroeconomic variable, the respective two explanatory variables and the eight principal components derived from the GDELT sentiment scores are included into the model.

The benchmark is an autoregressive model including the predicted indices and the explanatory variables as described above but excluding the GDELT sentiment factors, with the same lag as the above autoregressive framework.

The model is trained on 80% of the data and is tested out of sample using the most recent 20% of the data that have been set aside.

Performance is assessed using the root mean squared error (RMSE) on the test set predictions.

# 5. Research findings

This section presents the findings from the analysis set out in the previous section.

## 5.1. Granger causality test results

The sentiment scores from GDELT and the macroeconomic indices for 10 countries are tested for Granger causality, with a maximum lag of three months. Tables 2 and 3 display the number of BH-adjusted $p$-values that exhibit significance at 5% for each country's macroeconomic variable. The "Filtered" column refers to results from models including GDELT sentiment scores filtered for economic growth and inflation respectively, while the "Unfiltered" column shows the results for the models incorporating the unfiltered GDELT sentiment.

| Data set Country | Filtered | Unfiltered |
|---|---|---|
| US | 3 | 0 |
| UK | 28 | 0 |
| Germany | 8 | 7 |
| Norway | 30 | 0 |
| Poland | 5 | 0 |
| Turkey | 4 | 1 |
| Japan | 5 | 0 |
| South Korea | 6 | 0 |
| Brazil | 35 | 64 |
| Mexico | 12 | 0 |

Table 2: IP: Number of significant BH-adjusted $p$-values

| Data set Country | Filtered | Unfiltered |
|---|---|---|
| US | 14 | 0 |
| UK | 30 | 8 |
| Germany | 13 | 3 |
| Norway | 9 | 0 |
| Poland | 16 | 3 |
| Turkey | 57 | 1 |
| Japan | 18 | 0 |
| South Korea | 17 | 0 |
| Brazil | 41 | 0 |
| Mexico | 35 | 15 |

Table 3: CPI: Number of significant BH-adjusted $p$-values

Notwithstanding the limitations of the Granger causality test [30], the results show a pattern. For both macroeconomic variables, the filtered data sets exhibit consistent Granger causality across countries. The exception is Brazil in the case of IP, where unfiltered data set shows more Granger causality than the filtered one.

The analysis suggests that the filtering methodology introduced in section 3.2 adds value and is able to generate sentiment scores that have a relationship with economic indices.

## 5.2. Forecast error analysis

The respective filtered sentiment data sets are condensed into eight principal components using PCA. They are then used to predict IP and CPI, for 10 countries each with the model in Eq. (1). For comparison, the respective variables are forecast with unfiltered data. Tables 4 and 5 provide a high-level summary of the results from predictions on the most recent 20% of the data that has been set aside for for testing.

The "Filtered" column refers to results from models including filtered GDELT data, while the "Unfiltered" column refers to results for the models using unfiltered GDELT data. Blue (red) cells denote cases in which the models outperform (underperform) the benchmark. Numbers in parentheses correspond to the number of significant coefficients associated with GDELT factors in the model in Eq. (1), with the asterisks denoting the level of their statistical significance.

| Data set IP for | Filtered | Unfiltered |
|---|---|---|
| US | **(1), *(1) | ***(1), *(1) |
| UK |  |  |
| Germany |  |  |
| Norway | *(3) |  |
| Poland | *(1) | ***(1), **(2) |
| Turkey | **(1) | ***(1), **(2), *(4) |
| Japan | ***(1), **(1), *(2) |  |
| South Korea | **(3), *(2) |  |
| Brazil |  |  |
| Mexico | **(1) | **(1), *(1) |

Table 4: Results of the model in Eq. 1 applied to IP. Blue (red) cells denote cases in which the model outperforms (underperforms) the benchmark. Numbers in parentheses correspond to the number of significant coefficients associated with GDELT factors in the model in Eq. (1) ($***$ denotes at least one GDELT sentiment factor with $p$-value $< 0.01$, $** < 0.05$, $* < 0.1$).

Predicting IP, all models have a lag of one month except for Germany, which has lags of one and two months based on AIC and BIC. Nine out of ten country models using filtered data outperform the benchmark in terms of RMSE. Of the nine country models that perform better than the benchmark model, seven contain one or more statistically significant GDELT factors. Five of the outperforming models include a macroeconomic variable that is significant at 1, 5 or 10%. For those models using unfiltered sentiment data, two out of ten models outperform and one contains one or more significant GDELT factors. Both outperforming models include a statistically significant macroeconomic index.

| Data set<br>CPI for | Filtered | Unfiltered |
|---|---|---|
| US | **(1), *(2) | |
| UK | | ***(1), **(2) |
| Germany | **(1), *(1) | ***(1), **(1) |
| Norway | **(1), *(2) | *(1) |
| Poland | ***(2) | **(1), *(1) |
| Turkey | *(1) | ***(3), **(3), *(1) |
| Japan | ***(1), **(4), *(2) | **(1) |
| South Korea | **(1), *(1) | ***(2) |
| Brazil | **(1) | |
| Mexico | | |

Table 5: Results of the model in Eq. 1 applied to CPI. Blue (red) cells denote cases in which the model outperforms (underperforms) the benchmark. Numbers in parentheses correspond to the number of significant coefficients associated with GDELT factors in the model in Eq. (1) ($***$ denotes at least one GDELT sentiment factor with $p$-value $< 0.01$, $** < 0.05$, $* < 0.1$).

In the case of CPI, models for Germany, Japan and Mexico have lags of one month and two months, the rest of the models have a lag of one month, determined by evaluating AIC and BIC. Eight out of ten models using filtered sentiment data outperform their benchmark, seven with statistically significant GDELT factors and one or more significant macroeconomic indices. In contrast, three models using unfiltered data outperform; all including one or more significant sentiment factors as well as significant macroeconomic indices.

Results suggest that the filtering methodology isolates relevant signals, given that those models using filtered sentiment perform consistently better that those using unfiltered sentiment. Further, the findings indicate that sentiment scores extracted from GDELT improve predictions for IP and CPI for most countries.

### 5.3. Drivers of GDELT factors

In order to gain insights into the relationship between sentiment scores and the principal components derived from the filtered GDELT data, the loadings corresponding to each component are examined.

Loadings correspond to the strength of relationship between the original sentiment scores and the principal components, quantifying the relevance of the underlying sentiment scores in each of the components. They are derived by multiplying eigenvectors by the square root of the eigenvalues.

All sentiment scores from GDELT represent a specific emotion such as "cheerfulness", "euphoria" or "joy" and are manually mapped to seven universal emotions as set out by Ekman and Corduro [18]. These seven emotions define emotions as discrete, automatic reactions to events and stipulate that emotions such as happiness or anger describe groups of related states with distinct common traits. According to these seven groups, the above mentioned emotions are assigned to "happiness".

For each principal component, the squared loadings are summed according to these seven distinct emotions. Map-

ping sentiment scores onto emotions provides some interpretability to our analysis, by allowing us to investigate which emotions are associated with each principal component.
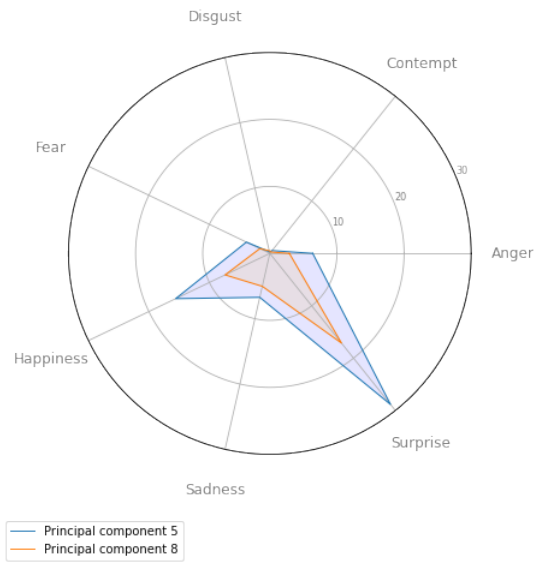


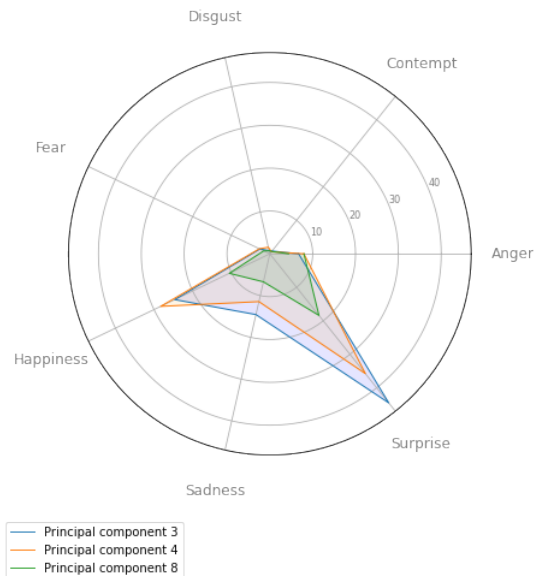**Figure 2:** IP: Significant principal components explained by emotions (US)



**Figure 3:** CPI: Significant principal components explained by emotions (US)

As an example, in Figs. 2 and 3 we show radar charts of the emotions associated to the loadings corresponding to the statistically significant principal components used to forecast IP and CPI in the US. As can be seen from Tables 4 and 5, the corresponding models outperform the benchmarks we considered and are associated with substantial statistical significance. The principal components explain 8.3% and 8.9% of the variance for the components shown in Figs. 2 and 3, respectively. Further charts can be provided upon request.

The findings from this example show that the factors we use to predict IP and CPI can be associated with well defined emotions. Therefore, movements in such emotions – as expressed in news articles published by global newspapers – contribute to explain movements in major macroeconomic indices. Of the seven distinct emotions, "surprise" and "happiness" have the strongest predictive power. This is the case across all principal components.

From a Keynesian point of view, these emotions can be considered "animal spirits" that guide human behaviour, ultimately being reflected in economic variables [29]. Likewise, Loewenstein argues that intense emotions – so-called visceral factors – affect the economy as they lead to individual decisions and actions and should therefore be included into economic models [36]. Our findings agree with the school of thought that human emotions have an impact on the economy.

## 6. Discussion

This study attempts to forecast macroeconomic indices using sentiment scores derived from GDELT. It introduces a filtering methodology to extract and aggregate large volumes of data. The methodology is applied to build data sets filtered for economic growth and inflation. The country-specific macroeconomic indices are forecast using data sets for IP and CPI, respectively, that take into account each country's trade links when applying location filters. The filtered data exhibits consistent Granger causality across the two macroeconomic variables, except for Brazil in the case of IP. Autoregressive models including the filtered data outperform their benchmark for most predicted variables and overall produce better results than those using unfiltered data. Mapping the GDELT sentiment scores onto distinct emotions helps understand how these emotions relate to each principal component and thus interpret our analysis, suggesting that "surprise" and "happiness" are their main drivers. The findings from this study agree with the school of thought that emotions drive human behaviour and eventually impact the economy.

### 6.1. Limitations and ideas for further research

This study examines linear relationships between GDELT variables and macroeconomic indices. Investigating non-linear interactions between these variables could potentially generate further insights and could be an extension to this experiment.

All GDELT sentiment scores are incorporated into the forecasting framework. An alternative approach could be to choose a subset of sentiment scores based on a feature selection criterion.

GDELT starts end of February 2015 and thus has a short track record. Particularly when modelling monthly data, the small amount of observations is likely to impact the significance of results.

## 7. Conclusions

Short-term forecasting using narrative and sentiment from media has emerged in recent years. The majority of research extracts data from anglophone sources, utilising means such as simple word count of positive and negative keywords; few studies use big data [43]. As a result, existing works do not capture the entire complexity of global news articles. This study expands the existing body of research by creating a data set that incorporates a wide array of emotions from newspapers around the world. To the best of our knowledge, the GCAM sentiment scores from the GDELT GKG have not yet been used to forecast macroeconomic variables; hence the experiment introduces a new data source.

The study represents a proof of concept showing that the filtering methodology presented captures relevant signals and that the data extracted from GDELT adds value when forecasting macroeconomic variables. The findings demonstrate that the sentiment factors derived from GDELT we use to predict IP and CPI can be linked to distinct emotions. Therefore, fluctuations in such emotions âĂŞ as expressed in news articles published by global newspapers âĂŞ help explain changes in major macroeconomic indices.

## References

[1] Allen, D.E., McAleer, M., Singh, A.K., 2019. Daily market news sentiment and stock prices. Applied Economics 51, 3212–3235.

[2] Ardia, D., Bluteau, K., Boudt, K., 2019. Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values. International Journal of Forecasting 35, 1370–1386.

[3] Baker, S., Bloom, N., Davis, S., Terry, S., 2020. Covid-induced economic uncertainty and its consequences. VoxEU. org 13.

[4] Baker, S.R., Bloom, N., Davis, S.J., 2016. Measuring economic policy uncertainty. The quarterly journal of economics 131, 1593–1636.

[5] Benjamini, Y., Yekutieli, D., 2005. False discovery rate–adjusted multiple confidence intervals for selected parameters. Journal of the American Statistical Association 100, 71–81.

[6] Bildirici, M.E., Kayıkçı, F., Onat, I.Ş., 2015. Baltic dry index as a major economic policy indicator: the relationship with economic growth. Procedia-Social and Behavioral Sciences 210, 416–424.

[7] Brooks, C., Tsolacos, S., 2010. Real Estate Modelling and Forecasting. doi:10.1017/CBO9780511814235.

[8] Brosch, T., Scherer, K.R., Grandjean, D.M., Sander, D., 2013. The impact of emotion on perception, attention, memory, and decision-making. Swiss medical weekly 143, w13786.

[9] Bruner, J.S., 1990. Acts of meaning. volume 3. Harvard university press.

[10] Buono, D., Kapetanios, G., Marcellino, M., Mazzi, G.L., Papailias, F., . Evaluation of nowcasting/flash estimation based on a big set of indicators. Paper prepared for the 16th Conference of IAOS .

[11] Buono, D., Mazzi, G.L., Kapetanios, G., Marcellino, M., Papailias, F., 2017. Big data types for macroeconomic nowcasting. Eurostat Review on National Accounts and Macroeconomic Indicators 1, 93–145.

[12] Casanova, C., Ortiz, A., Rodrigo, T., Xia, L., Iglesias, J., et al., 2017. Tracking chinese vulnerability in real time using Big Data. Technical Report.

[13] Chen, H.Y., Lo, T.C., 2019. Online search activities and investor attention on financial markets. Asia Pacific Management Review 24, 21–26.

[14] Clore, G.L., Palmer, J., 2009. Affective guidance of intelligent agents: How emotion controls cognition. Cognitive systems research 10, 21–30.

[15] Cruz, F.L., Troyano, J.A., Pontes, B., Ortega, F.J., 2014. Building layered, multilingual sentiment lexicons at synset and lemma levels. Expert Systems with Applications 41, 5984–5994.

[16] Datawheel, Simoes, A., Hidalgo, C.A., . The observatory of economic complexity. URL: https://oec.world/.

[17] Dodds, P.S., Harris, K.D., Kloumann, I.M., Bliss, C.A., Danforth, C.M., 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. PloS one 6, e26752.

[18] Ekman, P., Cordaro, D., 2011. What is meant by calling emotions basic. Emotion review 3, 364–370.

[19] Elshendy, M., Colladon, A.F., Battistoni, E., Gloor, P.A., 2018. Using four different online media sources to forecast the crude oil price. Journal of Information Science 44, 408–421.

[20] Elshendy, M., Fronzetti Colladon, A., 2017. Big data analysis of economic news: Hints to forecast macroeconomic indicators. International Journal of Engineering Business Management 9, 1847979017720040.

[21] Glaeser, E.L., Kim, H., Luca, M., 2017. Nowcasting the local economy: Using yelp data to measure economic activity. Technical Report. National Bureau of Economic Research.

[22] Granger, C.W., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica: journal of the Econometric Society , 424–438.

[23] Graves, A., Schmidhuber, J., 2005. Framewise phoneme classification with bidirectional lstm networks, in: Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005., IEEE. pp. 2047–2052.

[24] Hansen, S., McMahon, M., 2016. Shocking language: Understanding the macroeconomic effects of central bank communication. Journal of International Economics 99, S114–S133.

[25] He, H., Ma, Y., 2013. Imbalanced learning: foundations, algorithms, and applications. John Wiley & Sons.

[26] Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780.

[27] Hu, M., Liu, B., 2004. Mining and summarizing customer reviews, in: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 168–177.

[28] Kapetanios, G., Papailias, F., et al., 2018. Big data & macroeconomic nowcasting: Methodological review. Economic Statistics Centre of Excellence, National Institute of Economic and Social Research .

[29] Keynes, J.M., 2018. The general theory of employment, interest, and money. Springer.

[30] Leamer, E.E., 1985. Self-interpretation. Economics and Philosophy 1, 295âĂŞ302. doi:10.1017/S0266267100002546.

[31] Leetaru, K., 2016. Can we forecast conflict? A framework for forecasting global human societal behavior using latent narrative indicators. Ph.D. thesis. University of Illinois at Urbana-Champaign.

[32] Leetaru, K.H., a. The gdelt project. URL: https://www.gdeltproject.org/.

[33] Leetaru, K.H., b. THE GDELT GLOBAL KNOWLEDGE GRAPH (GKG) DATA FORMAT CODEBOOK v2.1. URL: http://gdeltproject.org/.

[34] Leetaru, K.H., 2015. Mining libraries: Lessons learned from 20 years of massive computing on the worldâĂŹs information. Information Services & Use 35, 31–50.

[35] Levenberg, A., Pulman, S., Moilanen, K., Simpson, E., Roberts, S., 2014. Predicting economic indicators from web text using sentiment composition. International Journal of Computer and Communication Engineering 3, 109–115.

[36] Loewenstein, G., 2000. Emotions in economic theory and economic behavior. American economic review 90, 426–432.

[37] Loughran, T., McDonald, B., 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. The Journal of Finance 66, 35–65.

[38] Magnini, B., Cavaglia, G., 2000. Integrating subject field codes into wordnet., in: LREC, pp. 1413–1418.

[39] Mihailov, A., Rumler, F., Scharler, J., 2011. The small open-economy new keynesian phillips curve: empirical evidence and implied inflation dynamics. Open Economies Review 22, 317–337.

[40] Nyman, R., Kapadia, S., Tuckett, D., Gregory, D., Ormerod, P., Smith, R., 2018. News and narratives in financial systems: exploiting big data for systemic risk assessment .

[41] Pekar, V., Binner, J., 2017. Forecasting consumer spending from purchase intentions expressed on social media, Association for Computational Linguistics.

[42] Piccardi, C., Tajoli, L., 2018. Complexity, centralization, and fragility in economic networks. PloS one 13, e0208265.

[43] Rousidis, D., Koukaras, P., Tjortjis, C., 2020. Social media prediction: a literature review. Multimedia Tools and Applications 79, 6279–6311.

[44] Salisu, A.A., Isah, K.O., Oyewole, O.J., Akanni, L.O., 2017. Modelling oil price-inflation nexus: The role of asymmetries. Energy 125, 97–106.

[45] Schaer, O., Kourentzes, N., Fildes, R., 2019. Demand forecasting with user-generated online information. International Journal of Forecasting 35, 197–212.

[46] Shiller, R.J., 2017. Narrative economics. American Economic Review 107, 967–1004.

[47] Slaper, T., Bianco, A., Lenz, P., 2018. Digital vapor trails: Using website behavior to nowcast entrepreneurial activity, in: 2nd International Conference on Advanced Reserach Methods and Analytics (CARMA 2018), Editorial Universitat Politècnica de València. pp. 107–113.

[48] Stern, S., Livan, G., Smith, R.E., 2020. A network perspective on intermedia agenda-setting. arXiv preprint arXiv:2002.05971 .

[49] Stock, J.H., Watson, M.W., 2002. Macroeconomic forecasting using diffusion indexes. Journal of Business & Economic Statistics 20, 147–162.

[50] Strapparava, C., Valitutti, A., et al., 2004. Wordnet affect: an affective extension of wordnet., in: Lrec, Citeseer. p. 40.

[51] Thorsrud, L.A., 2016. Nowcasting using news topics. big data versus big bank .

[52] Tuckett, D., Ormerod, P., Smith, R., Nyman, R., 2014. Bringing social-psychological variables into economic modelling: Uncertainty, animal spirits and the recovery from the great recession. Animal Spirits and the Recovery from the Great Recession (January 12, 2014) .

[53] Van Eyden, R., Difeto, M., Gupta, R., Wohar, M.E., 2019. Oil price volatility and economic growth: Evidence from advanced economies using more than a centuryâĂŹs data. Applied Energy 233, 612–621.